MACHINE LEARNING

EXERCISES

Elements of Bayesian decision theory

All the course material is available on the web site

Course web site: http://pralab.diee.unica.it/MachineLearning

PART 2: Elements of Bayesian decision theory

Required knowledge from Chapter 2 and problem-solving ability

Exercise 1

- Required knowledge: MAP decision rule, optimal Bayes decision threshold, error probability of a Bayesian pattern classifier, decision regions
- Problem-solving ability: you are able to apply MAP decision rule to simple one-dimensional classification problems, compute error probability for a simple Bayesian pattern classifier.

Exercise 2

- Required knowledge: MAP decision rule for multiple class cases (more than 2 classes), likelihood ratio test, error probability of a Bayesian pattern classifier for multiple class cases, decision regions, Gaussian distribution, erf function for error probability of Gaussian classifiers.
- Problem-solving ability: you are able to apply MAP decision rule to multiple class cases, compute error probability for multiple class cases.

Exercise 3

- Required knowledge: minimum risk decision rule, loss/cost matrix.
- Problem-solving ability: you are able to apply minimum risk decision rule to pattern classification, compute error probability taking into account costs of different classifications.

Exercise 3

- Required knowledge: minimum risk decision rule, loss/cost matrix, rejection option, Chow's reject rule.
- Problem-solving ability: you are able to apply minimum risk decision rule to pattern classification, also using rejection option and Chow's reject rule, compute error probability and compute reject probability.

Exercise 1

Let us suppose that we want to design an anti-spam filter based on the frequency (number of occurrences) of the word "viagra" in the e-mail body. Intuitively, a higher frequency of this keyword suggests a higher level of "spamminess" (level/degree of spam content) of the e-mail. Indeed, real anti-spam filters usually compute a "score" which is assigned to each incoming e-mail, based on the frequency of some keywords contained in the body message (e.g., the keyword "viagra").

1) Define a simple model (e.g., a linear one) for the probability density functions $p(x \mid \omega_{SPAM})$ and $p(x \mid \omega_{MAIL})$ of the classes "spam" and "ham".

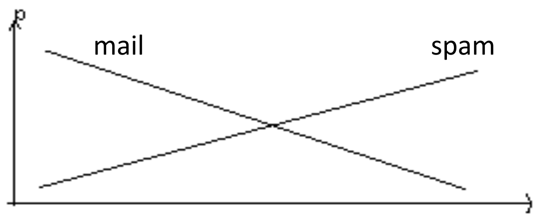
2) Find the Bayesian decision rule (let us indicate the Bayesian threshold with the letter x*) and compute the probability of error for the two cases of priors $P(\omega_{MAIL}) = P(\omega_{SPAM})$ and $P(\omega_{MAIL}) = 10\,P(\omega_{SPAM})$.

3) Plot the decision regions and the area under the probability density functions corresponding to the probability of error.

SOLUTION

1) Define a simple model (e.g. a linear model) for the functions $p(x \mid \omega_{SPAM})$ and $p(x \mid \omega_{MAIL})$

In the following figure, a possible, simple, linear model is shown.



x= frequency of "viagra"

Given that $\int p(x/\omega)\, dx = 1$, we have

$$p(x \mid \omega_{SPAM}) = \begin{cases} 2x & for \quad 0 < x < 1 \\ 0 & otherwise \end{cases}$$

$$p(x \mid \omega_{MAIL}) = \begin{cases} 2 - 2x & for \quad 0 < x < 1 \\ 0 & otherwise \end{cases}$$

x =1 is the maximum score.

2) Find the Bayesian decision rule (Bayesian threshold x*) and compute the probability of error for the two cases $P(\omega_{MAIL}) = P(\omega_{SPAM})$ and $P(\omega_{MAIL}) = 10\,P(\omega_{SPAM})$

MAP decision rule:

The mail is labelled as 'spam' if $P(\omega_{SPAM} / x) > P(\omega_{MAIL} / x)$

Using the Bayes decision rule: $P(\omega / x) = p(x / \omega)\dfrac{P(\omega)}{p(x)}$

We obtain the likelihood ratio $l(x) = \dfrac{p(x / \omega_{SPAM})}{p(x / \omega_{MAIL})} \underset{\omega_{MAIL}}{\overset{\omega_{SPAM}}{\gtrless}} \dfrac{P(\omega_{MAIL})}{P(\omega_{SPAM})}$

Bayes decision rule: if $x>x^*$ then e-mail is SPAM, otherwise is LEGAL (good) MAIL=HAM. About the terms "spam" and "ham", you find here some explanations:
http://www.todayifoundout.com/index.php/2010/09/how-the-word-spam-came-to-mean-junk-message/

If the two classes have **the same a priori probability** ($P(\omega_{MAIL})= P(\omega_{SPAM})$):

$$l(x) = \frac{p(x / \omega_{SPAM})}{p(x / \omega_{MAIL})} \underset{\omega_{MAIL}}{\overset{\omega_{SPAM}}{\gtrless}} 1 \Rightarrow \frac{2x}{2-2x} \underset{\omega_{MAIL}}{\overset{\omega_{SPAM}}{\gtrless}} 1 \Rightarrow x^* = \frac{1}{2}$$

The decision regions are:

MAIL: $x \in [0,\ x^*]$; SPAM: $x \in [x^*,\ 1]$; $x^* = \dfrac{1}{2}$

If the two classes have **different a priori probabilities** ($P(\omega_{MAIL})=10\ P(\omega_{SPAM})$), then

$P(\omega_{MAIL})=10/11$; $P(\omega_{SPAM})=1/11$;

$$l(x) = \frac{p(x / \omega_{SPAM})}{p(x / \omega_{MAIL})} \underset{\omega_{MAIL}}{\overset{\omega_{SPAM}}{\gtrless}} 10 \Rightarrow x^* = \frac{10}{11}$$

And the decision regions are

MAIL: $x \in [0,\ x^*]$; SPAM: $x \in [x^*,\ 1]$; $x^* = \dfrac{10}{11} \cong 0{,}909$

Note that the decision threshold for the case of priors $P(\omega_{MAIL})=10\ P(\omega_{SPAM})$ is much higher that the one for $P(\omega_{MAIL})=P(\omega_{SPAM})$. Is this reasonable? What is the meaning of this?

BAYESIAN ERROR:

(legitimate e-mails misclassified as spam) + (spam e-mails misclassified as legitimate)

$$P\{x \in R_{MAIL}, \omega_{SPAM}\} + P\{x \in R_{SPAM}, \omega_{MAIL}\} =$$
$$P\{x \in R_{MAIL} / \omega_{SPAM}\} P(\omega_{SPAM}) + P\{x \in R_{SPAM} / \omega_{MAIL}\} P(\omega_{MAIL})$$
$$\int_0^{x^*} p(x \mid \omega_{SPAM}) P(\omega_{SPAM}) dx + \int_{x^*}^{1} p(x \mid \omega_{MAIL}) P(\omega_{MAIL}) dx$$

If $P(\omega_{MAIL})=P(\omega_{SPAM})$,

$$err = \int_0^{1/2} 2x \cdot \frac{1}{2} dx + \int_{1/2}^{1} (2-2x) \cdot \frac{1}{2} dx = \frac{1}{4} = 1/8 + 1/8 = 25\%$$

Note that for this case with $P(\omega_{MAIL})= P(\omega_{SPAM})$ the two component of error are equal (1/8=0,125). The percentage of spam misclassified as ham and the percentage of ham misclassified as spam are the same (12,50%).

If $P(\omega_{MAIL})=10\ P(\omega_{SPAM})$

$$\int_0^{x^*} p(x \mid \omega_{SPAM}) P(\omega_{SPAM}) dx + \int_{x^*}^{1} p(x \mid \omega_{MAIL}) P(\omega_{MAIL}) dx =$$
$$= \int_0^{10/11} 2x \cdot \frac{1}{11} dx + \int_{10/11}^{1} (2-2x) \cdot \frac{10}{11} dx = 0.0751 + 0.00751 = \frac{110}{1331} \cong 0.0826 = 8.26\%$$

About 8 mails out of 100 will be misclassified. But note that:

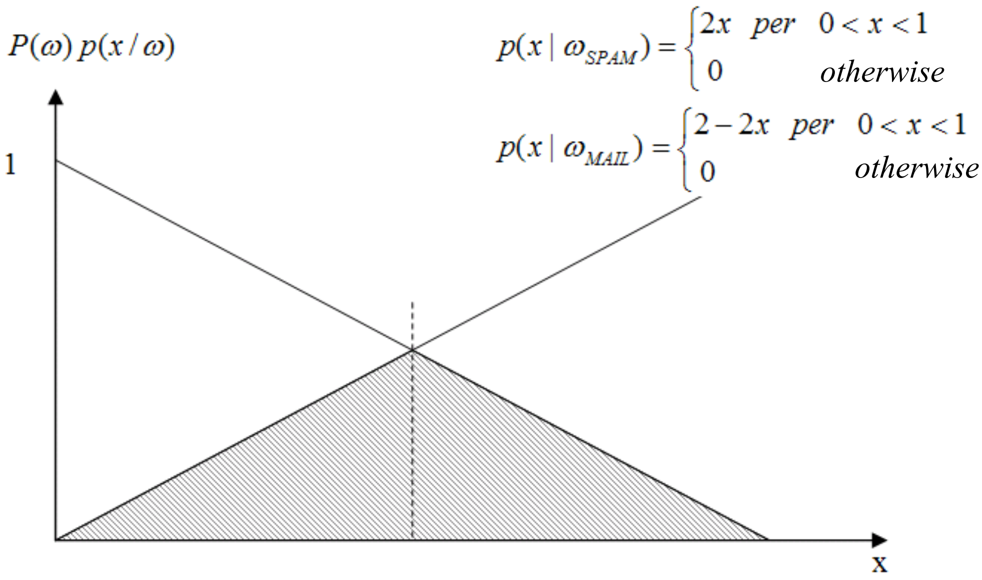
- the percentage of spam misclassified as ham is 7.51%
- the percentage of ham misclassified as spam is about 0.751%

This is because ham is much more likely.

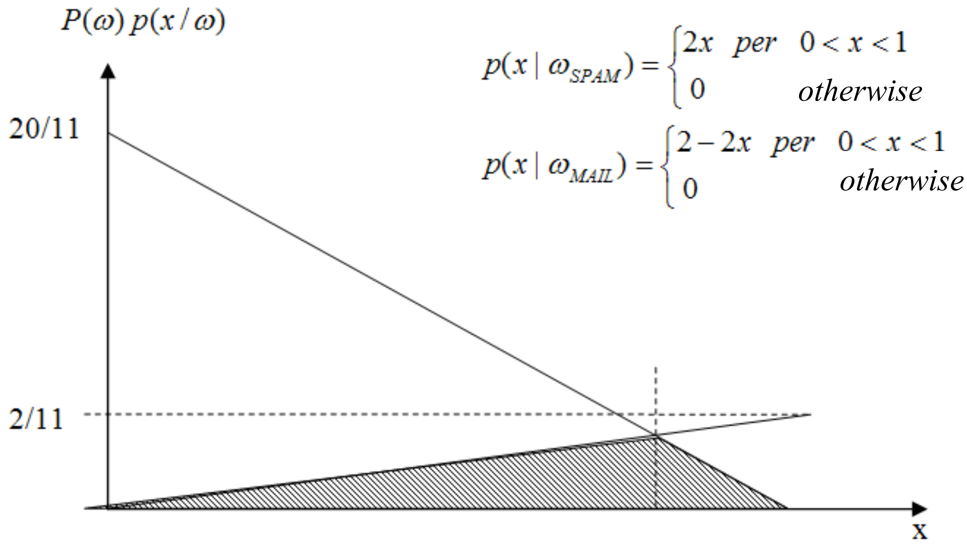
The total error is lower, but the error probability for spam misclassified as ham is higher.

3) Plot the decision regions and the area corresponding to the probability of error

If $P(\omega_{MAIL}) = P(\omega_{SPAM})$



If $P(\omega_{MAIL})=10\ P(\omega_{SPAM})$ => $P(\omega_{MAIL})=10/11$; $P(\omega_{SPAM})=1/11$;



Homework exercise: compute the probability of error and determine the decision regions for the case that the decision threshold is different from the Bayesian (optimal) one (i.e., x*). In particular, you should show that the probability of error is higher if the decision threshold is not the optimal one.

Exercise 2

Let us suppose that we want to recognise patterns belonging to three different kinds of network traffic[1]: normal traffic (class 1), and two different attacks (class 2 and 3). In this example, we use a single *feature* x (i.e., we have one-dimensional feature space), and we suppose that the network traffic has the following model:

$$P(\omega_1) = \frac{1}{2}; P(\omega_2) = P(\omega_3) = \frac{1}{4}$$

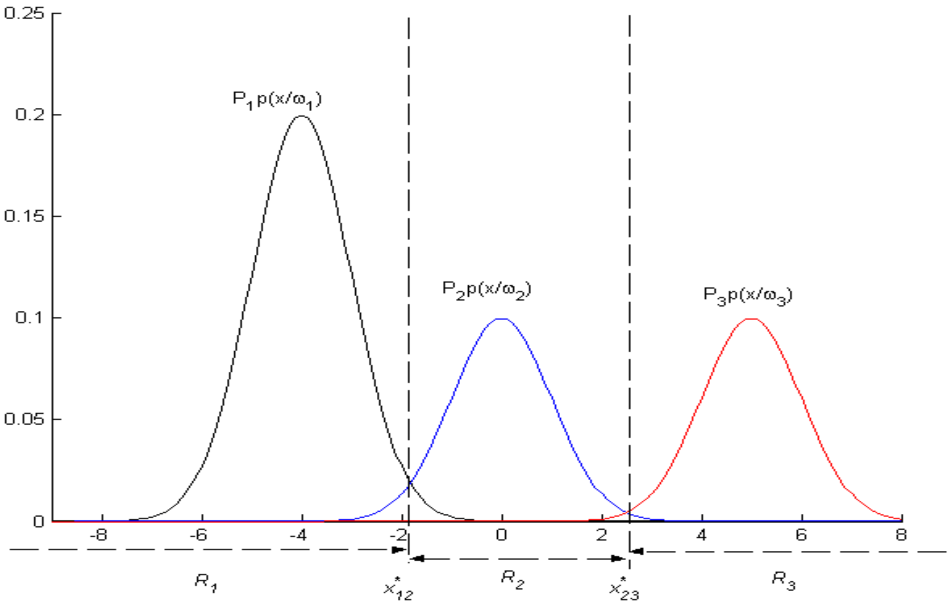
$$p(x/\omega_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma} \right)^2 \right];$$

$$\mu_1 = -4; \mu_2 = 0; \mu_3 = 5;\ \ \sigma_1 = \sigma_2 = \sigma_3 = 1;$$

a) Determine the decision regions using the likelihood ratio test;

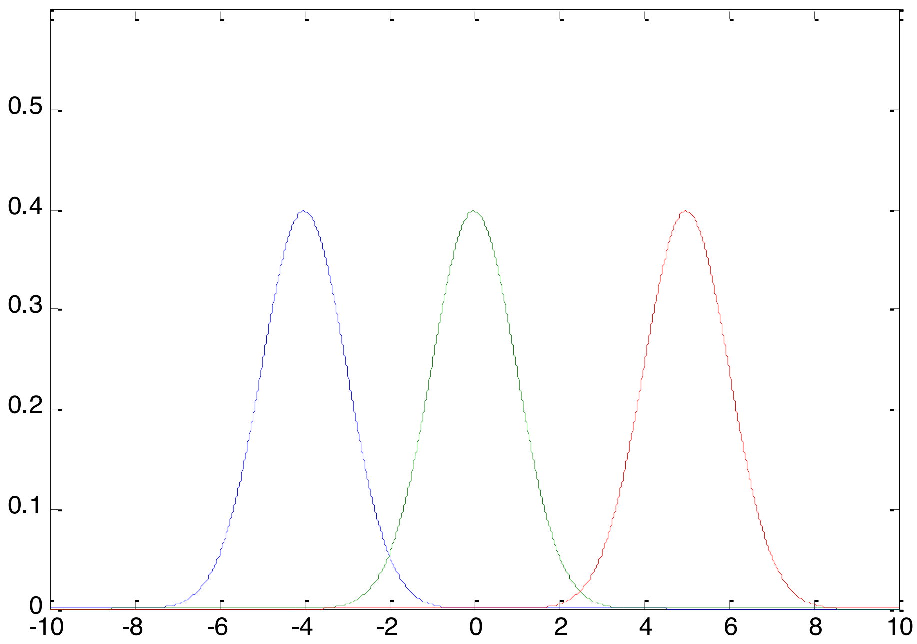
b) Compute the Bayesian error (minimum error probability).

The probability distributions and the decision regions can be represented graphically. The following figure plots the components $P_i\, p(x/\omega_i)$



[1] Part 2 of the course

while the following one is a plot of the three density functions $p(x/\omega_i)$



a) Determine the decision regions R_1, R_2, R_3

We can see immediately from the above figures that we **do not have** to consider the threshold between class 1 and class 3. This is clear for the simple one-dimensional example that we are considering. In the general case, we have to choose the thresholds **involving the two classes *s* and *t* such that**

$$P(\omega_s / x^*) = P(\omega_t / x^*) > P(\omega_i / x^*)$$

This is pretty much clear if you consider that the MAP rule selects the class with the maximum a posteriori probability (See Part 2 of the course). **You have a decision boundary if and only if the boundary is related to the maximum class probability, namely, it is the boundary between the two classes "s" and "t" that exhibit the maximum posterior probability:**

$$P(\omega_s / x^*) = P(\omega_t / x^*) > P(\omega_i / x^*)$$

We can determine explicitly the decision boundary between regions 1 and 2:

$$P_1\, p(x|\omega_1) = P_2\, p(x|\omega_2)$$

Let $x^*_{1,2}$ be the locus of points that satisfies the equation.

This value must be considered only if $P_1\, p\left(x^*_{1,2}\,|\omega_1\right) = P_2\, p\left(x^*_{1,2}\,|\omega_2\right) > P_3\, p(x^*_{1,2}\,|\omega_3)$

Otherwise, this threshold should not be considered. The same considerations apply to other boundaries.

Decision boundary between region 1 and 2:

$P_1\, p(x|\omega_1) = P_2\, p(x|\omega_2)$

$$\frac{1}{2}\exp\left[-\frac{1}{2}(x+4)^2\right] = \frac{1}{4}\exp\left[-\frac{1}{2}x^2\right]$$

$$\exp\left[-\frac{1}{2}(x+4)^2 + \frac{1}{2}x^2\right] = \frac{1}{2}$$

$$\exp[-8-4x] = \tfrac{1}{2}$$

$$x^*_{1,2} = -2 + \frac{1}{4}\ln(2) \cong -1.8267$$

In x^*_{12} we have $P_1\, p\left(x^*_{1,2}\,|\omega_1\right) = P_2\, p\left(x^*_{1,2}\,|\omega_2\right) > P_3\, p(x^*_{1,2}\,|\omega_3)$, in fact

$$P_1\, p\left(x^*_{1,2}\,|\omega_1\right) = \frac{1}{2}\cdot\frac{1}{\sqrt{2\,\pi}}\exp\left[-\frac{1}{2}\left(-2+\frac{1}{4}\ln(2)+4\right)^2\right] = \frac{1}{\sqrt{2\,\pi}}\cdot 0.0471$$

$$P_3\, p\left(x^*_{1,2}\,|\omega_3\right) = \frac{1}{4}\cdot\frac{1}{\sqrt{2\,\pi}}\exp\left[-\frac{1}{2}\left(-2+\frac{1}{4}\ln(2)-5\right)^2\right] = \frac{1}{\sqrt{2\,\pi}}\cdot 1.89\cdot 10^{-11}$$

so, the point x^*_{12} is a boundary between class 1 and 2.

Similarly, we can discriminate **between region R_2 and region R_3**

$P_3\, p(x|\omega_3) = P_2\, p(x|\omega_2)$

$$\frac{1}{4}\exp\left[-\frac{1}{2}(x-5)^2\right] = \frac{1}{4}\exp\left[-\frac{1}{2}x^2\right]$$

$$\exp\left[-\frac{1}{2}(x-5)^2 + \frac{1}{2}x^2\right] = 1$$

$$\exp\left[5x - \frac{25}{2}\right] = 1$$

$$x^*_{2,3} = \frac{5}{2}$$

In $x^*_{2,3}$ we have $P_3\, p\left(x^*_{2,3}\,|\omega_3\right) = P_2\, p\left(x^*_{2,3}|\omega_2\right) > P_1\, p(x^*_{2,3}\,|\omega_1)$, in fact

$$P_3\, p\left(x^*_{2,3}\,|\omega_3\right) = \frac{1}{4}\cdot\frac{1}{\sqrt{2\,\pi}}\exp\left[-\frac{1}{2}\left(\frac{5}{2}-5\right)^2\right] = \frac{1}{\sqrt{2\,\pi}}\cdot 0.110$$

$$P_1\, p(x^*_{2,3}|\omega_1) = \frac{1}{2}\cdot\frac{1}{\sqrt{2\,\pi}}\exp\left[-\frac{1}{2}\left(\frac{5}{2}+4\right)^2\right] = \frac{1}{\sqrt{2\,\pi}}\cdot 3.346\cdot 10^{-10}$$

so, the point $x^*_{2,3}$ is a boundary between class 2 and 3.

Similarly, we can discriminate **between region R_1 and region R_3**

$$P_1\, p(x|\omega_1) = P_3\, p(x|\omega_3)$$

$$\frac{1}{2}\exp\left[-\frac{1}{2}(x+4)^2\right] = \frac{1}{4}\exp\left[-\frac{1}{2}(x-5)^2\right]$$

$$\exp\left[-\frac{1}{2}(x+4)^2 + \frac{1}{2}(x-5)^2\right] = \frac{1}{2}$$

$$\exp\left[\frac{9}{2} - 9x\right] = \frac{1}{2}$$

$$x^*_{1,3} = \frac{1}{2} + \frac{1}{9}\ln(2) \cong 0.577$$

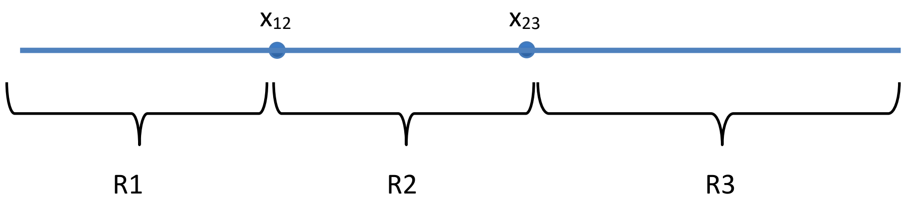
BUT in $x^*_{1,3}$ IT'S NOT SATISFIED that $P_1\, p(x^*_{1,3}\,|\omega_1) = P_3\, p(x^*_{1,3}|\omega_3) > P_2\, p(x^*_{1,3}\,|\omega_2)$, in fact

$$P_1\, p(x^*_{1,3}\,|\omega_1) = \frac{1}{2}\cdot\frac{1}{\sqrt{2\,\pi}}\exp\left[-\frac{1}{2}(0.577+4)^2\right] = \frac{1}{\sqrt{2\,\pi}}\cdot 1.4123\cdot 10^{-5}$$

$$P_2\, p(x^*_{1,3}\,|\omega_2) = \frac{1}{4}\cdot\frac{1}{\sqrt{2\,\pi}}\exp\left[-\frac{1}{2}(0.577)^2\right] = \frac{1}{\sqrt{2\,\pi}}\cdot 0.2117$$

so, the point $x^*_{1,3}$ **is NOT a boundary between class 1 and 3.**

Thus, we do not consider this threshold.



b) Compute the minimum Bayesian error

As we have seen in Chapter 2 of the course, it is simpler to compute first the probability of correct classification (classification accuracy) for a multiclass problem:

$$Pcorrect = \sum_{i=1}^{classes} P(x \in R_i, \omega_i) = \sum_{i=1}^{classes} P(x \in R_i / \omega_i) P(\omega_i) = \sum_{i=1}^{classes} P(\omega_i) \int_{R_i} p(x / \omega_i) dx;$$

And then, we can compute the error probability:

$$Perror = 1 - Pcorrect$$

$$P(correct) = P(\omega_1) \int_{-\infty}^{x_{12}^*} p(x / \omega_1) dx + P(\omega_2) \int_{x_{12}^*}^{x_{23}^*} p(x / \omega_2) dx +$$

$$+ P(\omega_3) \int_{x_{23}^*}^{\infty} p(x / \omega_3) dx = 0.493 + 0.24 + 0.248 = 0.981$$

$$P(error) = 1 - P(correct) = 0.019$$

In the following, we provide the complete calculation only for the first integral using the erf function (http://en.wikipedia.org/wiki/Error_function). Same calculations can be done for the other two integrals.

$$P_1 \int_{-\infty}^{x_{12}^*} p(x|\omega_1)\, dx = P1 \int_{-\infty}^{x_{12}^*} N(-4, 1)\, dx = P1 \int_{-\infty}^{x_{12}^*+4} N(0, 1)\, dx =$$

$$= P1 \cdot \left[\frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x_{12}^* + 4}{\sqrt{2}}\right)\right] = \frac{1}{2} \cdot \left[\frac{1}{2} + \frac{1}{2} \operatorname{erf}(1.5367)\right] \cong \frac{1}{2} \cdot \left[\frac{1}{2} + \frac{1}{2} \cdot 0.97\right] = 0.4925$$

Exercise 3

Let us suppose that we want to discriminate between normal and intrusive network traffic, namely, two data classes ω_N, normal traffic, and ω_{INT}, intrusive network traffic. We suppose to use a single *feature* x to characterize traffic data (one-dimensional feature space), and we assume that the model of the network traffic is the following:

$$P(\omega_N) = \frac{1}{2}; P(\omega_{INTR}) = \frac{1}{2}$$

$$p(x/\omega_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2\right];$$

$$\mu_N = 0; \mu_{INTR} = 4; \;\; \sigma_N = \sigma_{INTR} = 1;$$

Let the cost of missing the detection of intrusion be ten times higher than the opposite error (a normal traffic is wrongly recognized as an intrusion).

a) Determine the decision regions using the likelihood ratio, <u>without</u> considering the costs of errors.

b) Specify the loss (cost) matrix that satisfies the above assumption.

c) Determine the decision regions that minimize the risk, and compute the related classification error.

Solution

a) Determine the decision regions using the likelihood ratio, <u>without</u> considering the cost of misclassification.

$$l(x) = \frac{p(x/\omega_N)}{p(x/\omega_{INTR})} \underset{\omega_{INTR}}{\overset{\omega_N}{\gtrless}} \frac{P(\omega_{INTR})}{P(\omega_N)} = \theta;$$

We can determine explicitly the decision regions:

$$\theta = \frac{P(\omega_{INTR})}{P(\omega_N)} = 1$$

$$l(x) = \frac{p(x/\omega_N)}{p(x/\omega_{INTR})} = \exp\left[\frac{1}{2}\left(\left(\frac{x-\mu_{INTR}}{\sigma}\right)^2 - \left(\frac{x-\mu_N}{\sigma}\right)^2\right)\right] = \exp(8-4x)$$

Detailed calculations:

$$l(x) = exp\left[\frac{1}{2}\left(\left(\frac{x-4}{1}\right)^2 - \left(\frac{x-0}{1}\right)^2\right)\right] = exp\left[\frac{1}{2}(x^2 + 16 - 8x - x^2)\right] =$$

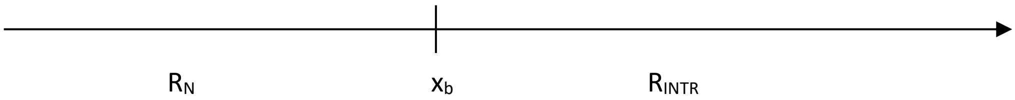
$$= exp[8-4x]$$

$$l(x) = \theta \Rightarrow exp[8-4x] = 1 \Rightarrow 8-4x = \ln(1) \Rightarrow$$

$$x_b = 2$$

$$l(x) > \theta \Rightarrow x < x_b$$

Let R_N e R_{INTR} be the two decision regions. If *x* belongs to R_N the traffic is labelled as 'normal'. If *x* belongs to R_{INTR} the incoming traffic is labelled as 'intrusion'.



BAYESIAN ERROR:

Two components of error probability: (intrusions wrongly labeled as normal traffic) + (normal traffic wrongly labeled as intrusion)

(the actual value of the integrals can be found by looking at the table of values of the erf function)

$$P\{x \in R_N, x \in \omega_{INTR}\} + P\{x \in R_{INTR}, x \in \omega_N\} =$$

$$P\{x \in R_N / \omega_{INTR}\} P(\omega_{INTR}) + P\{x \in R_{INTR} / \omega_N\} P(\omega_N) =$$

$$\int_{-\infty}^{x^*} p(x \mid \omega_{INTR}) P(\omega_{INTR}) dx + \int_{x^*}^{\infty} p(x \mid \omega_N) P(\omega_N) dx =$$

$$\frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{2} \exp\left[-\frac{1}{2}(x-4)^2\right] dx + \int_{2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x)^2\right] dx\right] =$$

$$\frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2} \exp\left[-\frac{1}{2}(y)^2\right] dy + \int_{2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x)^2\right] dx\right] =$$

$$\frac{1}{2}[0.0228 + 0.0228] = 0.0228$$

b) Specify the cost matrix.

Let the cost of missing the detection of intrusion be ten times higher than the opposite error.

We can indicate with $\lambda_{N,Intr}$ the cost when the traffic is *intrusive* but it is classified as *normal* (and vice versa for $\lambda_{Intr,N}$). Therefore, we can set $\lambda_{N,Intr} = 10 \cdot \lambda_{Intr,N}$. A possible solution is

$$\Lambda = \begin{bmatrix} \lambda_{N,N} & \lambda_{N,Intr} \\ \lambda_{Intr,N} & \lambda_{Intr,Intr} \end{bmatrix} = \begin{bmatrix} 0 & \lambda_{N,Intr} \\ \lambda_{Intr,N} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 10 \\ 1 & 0 \end{bmatrix}$$

Note: the decision regions that minimize the risk are defined by the ratio between costs, so the individual values do not matter.

c) Determine the decision regions that minimize the risk, and compute the classification error.

We must minimise the following risk terms:

$$R(\omega_N/\mathbf{x}) = \lambda_{N,N}P(\omega_N/\mathbf{x}) + \lambda_{N,INTR}P(\omega_{INTR}/\mathbf{x})$$
$$R(\omega_{INTR}/\mathbf{x}) = \lambda_{INTR,N}P(\omega_N/\mathbf{x}) + \lambda_{INTR,INTR}P(\omega_{INTR}/\mathbf{x})$$

According to the minimum risk criterion:

The traffic data will be assign to class ω_N if $R(\omega_N/\mathbf{x}) < R(\omega_{Intr}/\mathbf{x})$, it will be assign to ω_{INTR} otherwise.

Therefore, we assign the patter to ω_N if

$$R(\omega_N/\mathbf{x}) < R(\omega_{Intr}/\mathbf{x}) \rightarrow \lambda_{N,INTR}P(\omega_{INTR}/\mathbf{x}) < \lambda_{Intr,N}P(\omega_N/\mathbf{x})$$

Using the Bayes theorem:

$$\lambda_{N,INTR}\, p(x/\omega_{INTR})P(\omega_{INTR}) < \lambda_{Intr,N}\, p(x/\omega_N)P(\omega_N)$$

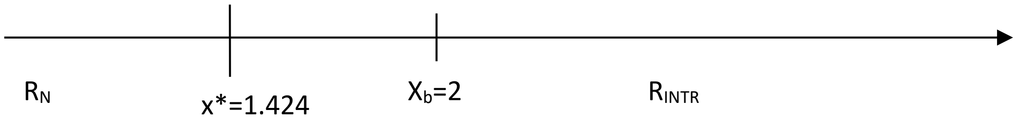
$$\frac{p(x/\omega_N)}{p(x/\omega_{INTR})} > \frac{\lambda_{N,INTR}}{\lambda_{Intr,N}} \frac{P(\omega_{INTR})}{P(\omega_N)} \rightarrow$$

$$l(x) = \frac{p(x/\omega_N)}{p(x/\omega_{INTR})} = \exp\left[\frac{1}{2}\left(\left(\frac{x-\mu_{INTR}}{\sigma}\right)^2 - \left(\frac{x-\mu_N}{\sigma}\right)^2\right)\right] = \exp(8-4x)$$

$$\frac{p(x/\omega_N)}{p(x/\omega_{INTR})} > 10\frac{P(\omega_{INTR})}{P(\omega_N)}$$

If we apply the logarithm function to the above equation:

$$l(x) = \theta \rightarrow 8 - 4x = \ln(10) \rightarrow x^* = \frac{1}{4}\left(8 - \ln(10)\right) \cong 1.424$$



Note that the decision region R_{INTR} is larger now, due to the fact that the cost of missing the detection of intrusion is ten times higher than the opposite error (a normal traffic is wrongly recognized as an intrusion). Accordingly, we prefer to make a decision for the class "intrusion" at the cost of a higher rate of false detections (normal traffic wrongly recognized as an intrusion)

ERROR PROBABILITY: (intrusion labeled as normal traffic) + (normal traffic labeled as intrusion)

$P\{x \in R_N, x \in \omega_{Intr}\} + P\{x \in R_{Intr}, x \in \omega_N\} =$

$= P(\omega_{Intr}) \cdot P\{x \in R_N | \omega_{Intr}\} + P(\omega_N) \cdot P\{x \in R_{Intr} | \omega_N\} =$

$$= P(\omega_{Intr}) \cdot \int_{-\infty}^{x^*} p(x|\omega_{Intr})\, dx \; + P(\omega_N) \cdot \int_{x^*}^{\infty} p(x|\omega_N)\, dx =$$

$$= P(\omega_{Intr}) \cdot \int_{-\infty}^{x^*} N(4,1)\, dx \; + P(\omega_N) \cdot \int_{x^*}^{\infty} N(0,1)\, dx =$$

$= 0.0025 + 0.0386 = 0.0411$

Note that it was 0.0228 using the Bayesian threshold! Now it's higher.

The cost of missing the detection of intrusion is ten times higher than the opposite error. This choice minimizes the error related to miss the detection:

$P\{x \in R_N, x \in \omega_{Intr}\} = 0.0025$

but the total error is greater than the Bayesian error due to the increase of the error related to the normal traffic labelled as intrusive:

$P\{x \in R_{Intr}, x \in \omega_N\} = 0.0386$

Exercise 4

Let us suppose that we want to diagnose a disease of which we know the prior probability:

$P(\omega_{sane=healthy})=0.85$, $P(\omega_{AFFECTED})=0.15$

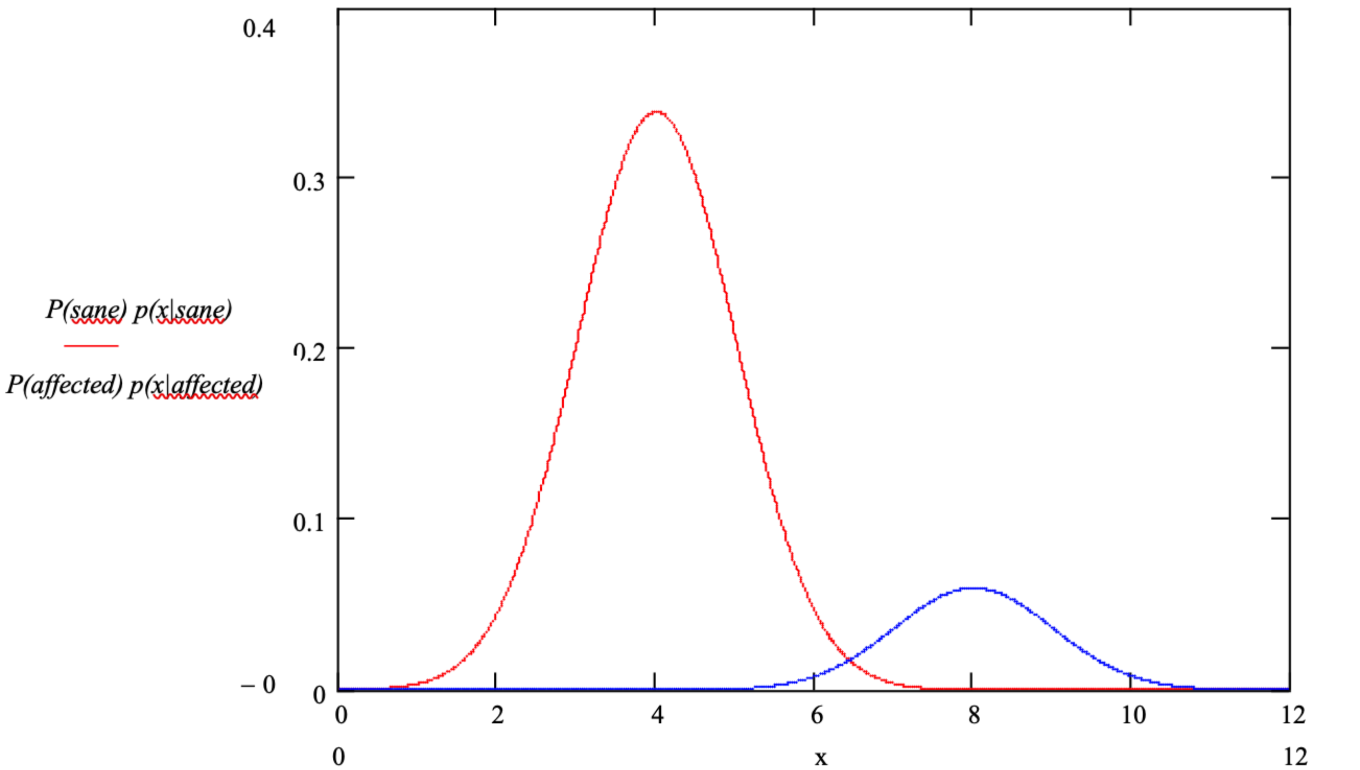
$P(\omega_{AFFECTED})$ is the prior probability that a person within a given population is affected by this disease.

Sane=Healthy.

The disease can be diagnosed by the amount of a certain substance in the blood. The amount of this substance is higher for people affected by the disease.

Let $\mu_s=4$ and $\mu_a=8$ be the average amount of this substance, respectively, for people *not affected* (**healthy** people) and *affected* by the disease. The amount of substance in the two cases is Gaussian distributed around the average value, with $\sigma=1$

$p(x|\omega_i) = N(\mu_i, \sigma^2)$; i=1 healthy, i=2 affected



Questions

a) How can we achieve the minimum classification error? Compute separately the two error probabilities for the two classes $\omega_{sane=healthy}$ and $\omega_{AFFECTED}$.
b) Let us suppose that we need to reduce the number of false negatives (affected people misclassified as healthy people), even at the cost of increasing the number of false positives (false alarms). What should we do? which technique that we have seen in Chapter 2 we should use?
c) Can we further reduce the error probability? Using which technique that we have seen in Chapter 2?

a) How can we achieve the minimum classification error? Compute separately the two error probabilities for the two classes.

Using the likelihood ratio, we can obtain the minimum error probability:

$$l(x) = \frac{p(x/\omega_{AFFECTED})}{p(x/\omega_{SANE})} \underset{\omega_{SANE}}{\overset{\omega_{AFFECTED}}{\gtrless}} \frac{p(x/\omega_{SANE})}{p(x/\omega_{AFFECTED})} = \theta$$

$$l(x) = \frac{p(x/\omega_{AFFECTED})}{p(x/\omega_{SANE})} \underset{\omega_{SANE}}{\overset{\omega_{AFFECTED}}{\gtrless}} \frac{P(\omega_{SANE})}{P(\omega_{AFFECTED})} = \theta$$

$$l(x) = \frac{\exp\left[-\frac{1}{2}(x-8)^2\right]}{\exp\left[-\frac{1}{2}(x-4)^2\right]} = \exp\left[\frac{1}{2}(x^2 - 8x + 16\ ) - \frac{1}{2}(x^2 - 16x + 64)\right] = \exp[4x - 24]$$

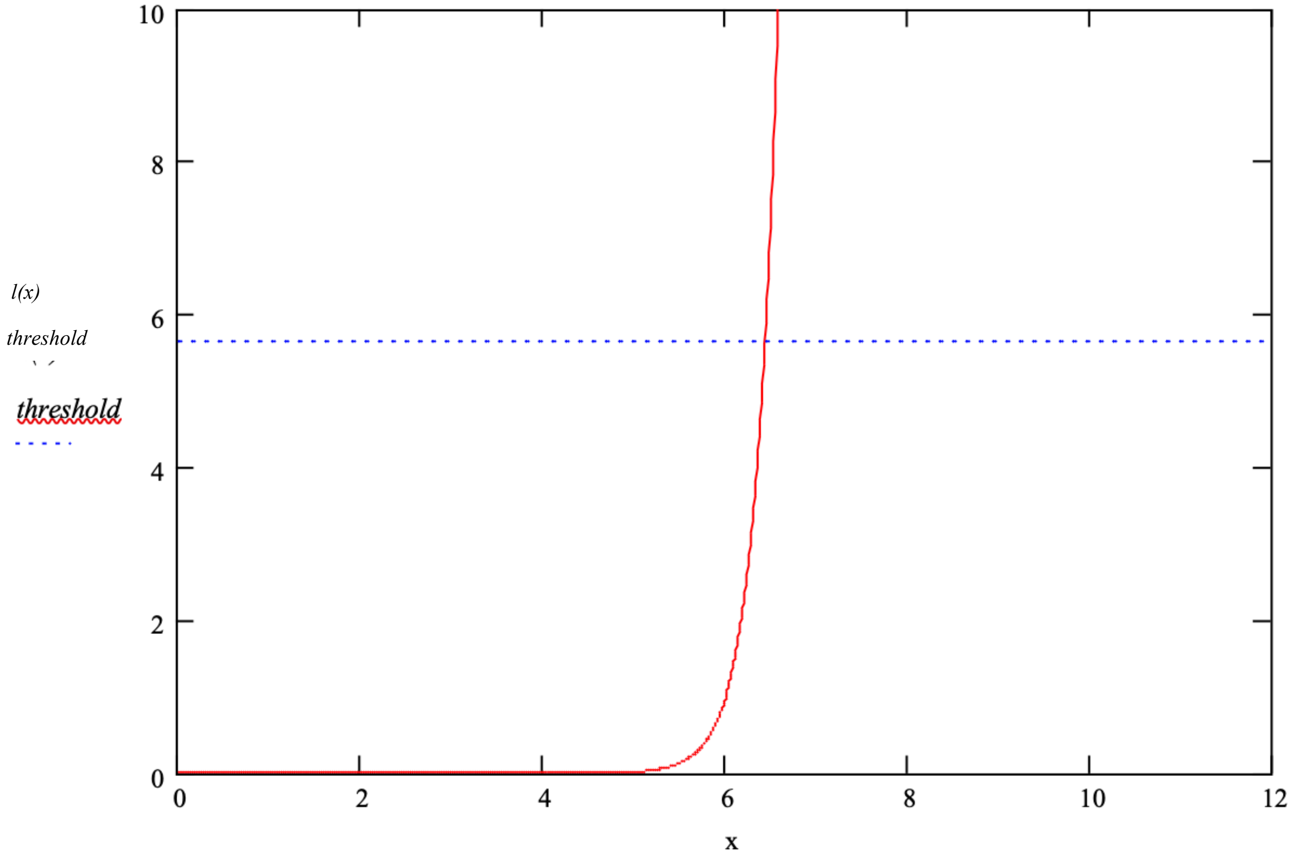
$$\theta = \frac{85}{15}$$

We classify a person as belonging to Class *Affected* or Class *Healthy* according to the decision rule:

$$\exp[4x - 24] \underset{\omega_S}{\overset{\omega_A}{\gtrless}} \frac{85}{15};$$

$$x \underset{\omega_S}{\overset{\omega_A}{\gtrless}} 6 + \frac{1}{4}\ln\left(\frac{85}{15}\right); \qquad x \underset{\omega_S}{\overset{\omega_A}{\gtrless}} 6.4337\ ; \qquad x_b = 6.4337$$

If $x<x_b$, the patient is labelled as HEALTHY (“sane”), otherwise as AFFECTED.



Note that in the above figure we depict the likelihood function:

$$(x) = \frac{\exp\left[-\frac{1}{2}(x-8)^2\right]}{\exp\left[-\frac{1}{2}(x-4)^2\right]} = \exp\left[\frac{1}{2}(x^2-8x+16\,) - \frac{1}{2}(x^2-16x+64)\right] = \exp[4x-24]$$

The blue line is the “threshold” $\theta = \frac{85}{15}$=5,66

The Bayesian error probability is:

$$\int_{-\infty}^{x_B} p(x/\omega_{AFFECTED})\, P(\omega_{AFFECTED})\, dx + \int_{xB}^{\infty} p(x/\omega_{SANE})\, P(\omega_{SANE})\, dx \cong 15.14\ \times 10^{-3}$$

This is the minimum error. It means that 15 patients every 1000 will be misdiagnosed.

In particular, the false positive error rate (HEALTHY patients misclassified as AFFECTED) is

$$\int_{x_B}^{\infty} p(x/\omega_{SANE})\, P(\omega_{SANE})\, dx \cong 6.35\ \times 10^{-3}$$

Whereas the false negative error rate (sick patients not detected as AFFECTED) is

$$\int_{-\infty}^{x_B} p(x/\omega_{AFFECTED})\, P(\omega_{AFFECTED})\, dx \cong 8.79\ \times 10^{-3}$$

b) Let us suppose that we need to reduce the number of false negatives (sick people misclassified as healthy people), even at the cost of incrementing the number of false positives (false alarms). What should we do?

We can introduce the costs λ_{ij} for a two-class problem (see Part 2 of the course, minimum risk decision rule), so that the pattern is classified as belonging to ω_1 if :

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\frac{P(\omega\lambda_2)}{P(\omega_1)}$$

That becomes $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12}}{\lambda_{21}}\frac{P(\omega_2)}{P(\omega_1)}$ if the cost of a correct decision is zero.

Note that the likelihood ratio will now be compared with a threshold that depends on the cost of the errors. We can associate a higher cost to the error that we want to reduce.

Following the above inequality, the patient will be classified as AFFECTED if:

$$\frac{p(x|\omega_S)}{p(x|\omega_S)} > \frac{\lambda_{AS}}{\lambda_{SA}}\frac{P(\omega_S)}{P(\omega_A)} = \theta$$

where

λ_{SA} is the cost of a false negative (affected people classified as healthy)

λ_{AS} is the cost of a false positive (healthy people classified as affected)

Introducing the following cost matrix:

$$\Lambda = \begin{pmatrix} 0 & 1 \\ 10 & 0 \end{pmatrix} = \begin{pmatrix} \lambda_{AA} & \lambda_{AS} \\ \lambda_{SA} & \lambda_{SS} \end{pmatrix};$$

The patient is classified as AFFECTED if:

$$l(x) = \frac{p(x|\omega_M)}{p(x|\omega_S)} > \frac{1}{10}\frac{P(\omega_S)}{P(\omega_A)} = \frac{1}{10}\frac{85}{15} = \theta$$

The minimum risk criterion is

$l(x) = \exp[4x - 24]$ as in the previous case

$$\theta = \frac{1}{10}\frac{85}{15}$$

$$\exp[4x - 24] \underset{\omega_S}{\overset{\omega_A}{\gtrless}} \frac{1}{10}\frac{85}{15}; \qquad x \underset{\omega_S}{\overset{\omega_A}{\gtrless}} 6 + \frac{1}{4}\ln\left(\frac{1}{10}\frac{85}{15}\right);$$

$x \underset{\omega_S}{\overset{\omega_A}{\gtrless}} 5.858$; $x_r = 5.858$ (Note that without considering the costs it was $x_b = 6.4337$)

The decision threshold is now lower. Thus, the amount of substance required to classify a patient as AFFECTED is lower.

The error is now

$$\int_{-\infty}^{x_r} p(x/\omega_{AFFECTED})\, P(\omega_{AFFECTED})\, dx + \int_{x_r}^{\infty} p(x/\omega_{SANE})\, P(\omega_{SANE})\, dx \cong 29.26 \times 10^{-3}$$

(the Bayesian error was 15.14 10^{-3})

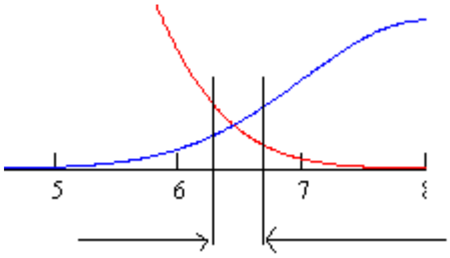
FP (false positives): 26.84 10^{-3} (it was 6.35 10^{-3})

FN (false negatives): 2.41 10^{-3} (it was 8.79 10^{-3})

We have reduced the number of False Negatives, at the expense of a higher overall classification error.

c)Can we further reduce the error probability? Using which technique that we have seen in Chapter 2?

Yes, if we can omit decisions in some cases, that is, we can use the “reject” option that we have seen in Chapter 2. For example, the pattern classification machine could not make any decision for some patients and they could be asked to repeat the blood test later, or we can decide to perform a more intrusive examination (for example, a biopsy).



If we refuse to classify (*reject option*) patterns that lay within the “reject region” shown in the figure, we can reduce the classification error, at the expense of requiring a further classification step to deal with the rejected patterns.

We have thus two classes (HEALTHY, AFFECTED), and three possible actions (REJECT, classify as HEALTHY, classify as AFFECTED)

We know that the Bayesian threshold is x_b= 6.434, that provides:

Overall classification error=15.14 10^{-3}

False positives=6.35 10^{-3}

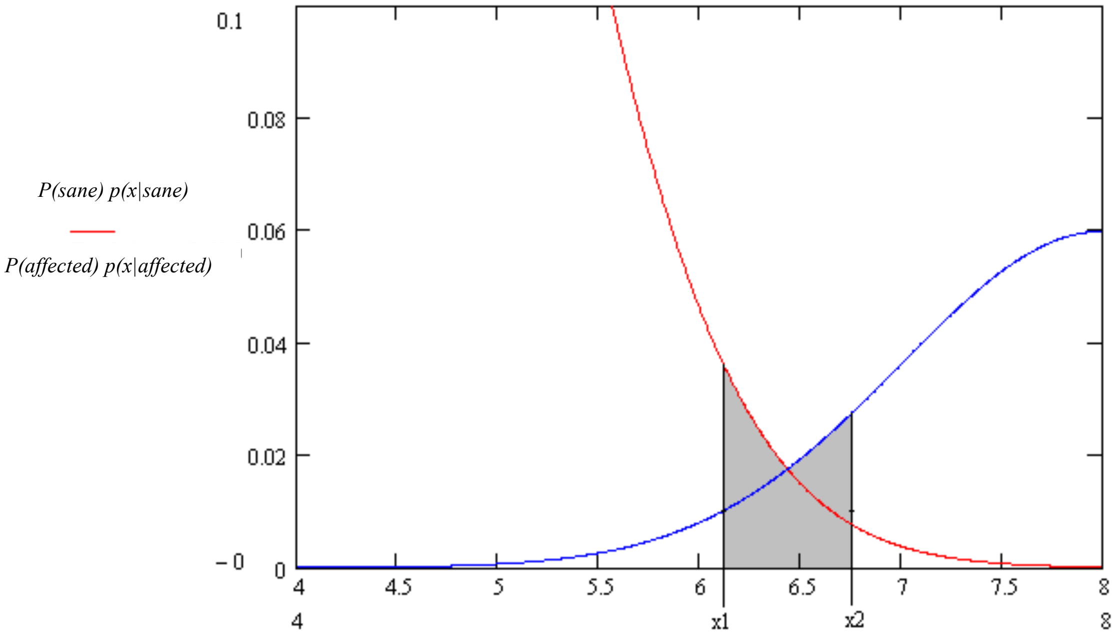
False negatives= 8.79 10^{-3}

As a first attempt, we could define a reject region [x1, x2] with

$x1=[x_b-(x_b/20)]=6.112$; $x2=[x_b+(x_b/20)]=6.755$;

Note that this is only an EXAMPLE; it is NOT the optimal choice! Just an attempt to check what happens.

In this way we reject patterns that could lead to misclassification.



False positive rate (it was 6.35 10^{-3} without rejection):

$$\int_{x_2}^{\infty} p(x/sane)\, P(\omega_{SANE})\, dx \cong 2.492 \times 10^{-3}$$

False negative rate (was 8.79 10^{-3} without rejection)

$$\int_{-\infty}^{x_1} p(x/\omega_{AFFECTED})\, P(\omega_{AFFECTED})\, dx \cong 4.43 \times 10^{-3}$$

Now we have greatly reduced the error. However, **we are omitting decisions (rejecting) for about 24 patients every 1000:**

$$\int_{x_1}^{x_2} p(x/\omega_{AFFECTED})\, P(\omega_{AFFECTED})\, dx + \int_{x_1}^{x_2} p(x/\omega_{SANE})\, P(\omega_{SANE})\, dx \cong 23.82 \times 10^{-3}$$

A principled criterion to choose the rejection threshold is to introduce a cost for the reject option:

Example 1 – a wrong choice of the cost matrix

$$\Lambda = \begin{pmatrix} \lambda_R = \lambda_{R1} & \lambda_R = \lambda_{R2} \\ \lambda_{AA} & \lambda_{AS} \\ \lambda_{SA} & \lambda_{SS} \end{pmatrix} = \begin{pmatrix} \lambda_R & \lambda_R \\ \lambda_C & \lambda_E \\ \lambda_E & \lambda_C \end{pmatrix} = \begin{pmatrix} 0.96 & 0.96 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The cost of each action is:

$$\lambda = \begin{cases} \lambda_C = 0, Correct\ action \\ \lambda_R = 0.96, reject \\ \lambda_E = 1, error \end{cases}$$

We can evaluate the *risk*:

Classification for class ω1

$$R(classification = \omega_1/x) = \lambda_E[P(\omega_2/x)] + \lambda_C[P(\omega_1/x)] =$$

$$= \lambda_E[P(\omega_2/x)] = [P(\omega_2/x)] = [1 - P(\omega_1/x)]$$

Classification for class ω2

$$R(classification = \omega_2/x) = \lambda_E[P(\omega_1/x)] + \lambda_C[P(\omega_2/x)] =$$

$$= \lambda_E[P(\omega_1/x)] = [P(\omega_1/x)] = [1 - P(\omega_2/x)]$$

The general formulation for class ω_i (if $\lambda_c = 0$) is

$$R(choice = \omega_i/x) = \sum_{\substack{j, \\ j \neq i}} \lambda_E\left[P\left(\omega_j/x\right)\right] + \lambda_C[P(\omega_i/x)] = \lambda_E[1 - P(\omega_i/x)]$$

We minimise the risk if we make a decision for the class ω_i according to the following rule:

$R(choice = \omega_i|x) < R\left(choice = \omega_j|x\right)$

$\lambda_E[1 - P(\omega_i|x)] < \lambda_E\left[1 - P\left(\omega_j|x\right)\right]$

$P(\omega_i|x) > P\left(\omega_j|x\right)$ $, \forall j \neq i$

The risk of the reject option is:

$$R(reject/x) = \sum_i \lambda_R[P(\omega_i/x)] = \lambda_R$$

Therefore, **we CLASSIFY the pattern if** the cost of rejection is less than the cost of the classification, that is:

$\min[R(\omega_i|x)] < R(reject|x) \Rightarrow$

$\Rightarrow \min\big[\lambda_E[1 - P(\omega_i|x)]\big] < \lambda_R \Rightarrow \lambda_E \Rightarrow \lambda_E - \lambda_E \max[P(\omega_i\,|x)] < \lambda_R \Rightarrow$

$\Rightarrow \max[P(\omega_i\,|x)] > 1 - \frac{\lambda_R}{\lambda_E}$

And we **REJECT the pattern if**

$\max[P(\omega\_i\,|x)] < 1 - \frac{\lambda_R}{\lambda_E}$

In other words, we assign the pattern to the class ω_i if

i) $R(\omega_i|x) < R(\omega_j|x), \forall j \neq i$

ii) $R(choice = \omega_i|x) < R(reject|x)$

that is

i) $P(\omega_i|x) > P(\omega_j|x), \forall j \neq i$

ii) $\lambda_E[1 - P(\omega_i|x)] < \lambda_R \Longrightarrow P(\omega_i|x) > 1 - \frac{\lambda_R}{\lambda_E}$

If only (i) is satisfied, then we reject the pattern.

Note that with the **wrong choice** of the cost matrix:

$$\Lambda = \begin{pmatrix} \lambda_R = \lambda_{R1} & \lambda_R = \lambda_{R2} \\ \lambda_{AA} & \lambda_{AS} \\ \lambda_{SA} & \lambda_{SS} \end{pmatrix} = \begin{pmatrix} \lambda_R & \lambda_R \\ \lambda_C & \lambda_E \\ \lambda_E & \lambda_C \end{pmatrix} = \begin{pmatrix} 0.96 & 0.96 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

the condition: ii)$\lambda_E[1 - P(\omega_i|x)] < \lambda_R \Longrightarrow P(\omega_i|x) > 1 - \frac{\lambda_R}{\lambda_E}$

is equal to: $[1 - P(\omega_i|x)] < 0.96 \Longrightarrow P(\omega_i|x) > 1 - \frac{0.96}{1} = 0.04$

that is always true. **Therefore, we never reject a pattern**.

This is due to the high value of the reject cost (λ_R =0.96).

We obtain the same result with the CHOW's rule.

According to the Chow's rule, the reject threshold *T* (see Chapter 2 of the course) is equal to:

$$T = \frac{\lambda_E - \lambda_R}{\lambda_E - \lambda_C}$$

According to the Chow's rule, we classify a pattern as belong to the class ω_i if $\max[P(\omega_i \mid x)] > T$, we reject it otherwise.

In our example, we have:

$$\lambda = \begin{cases} \lambda_C = 0, Correct\ class \\ \lambda_C = 0.96, reject \\ \lambda_C = 1, error \end{cases}$$

$$T = \frac{\lambda_E - \lambda_R}{\lambda_E - \lambda_C}$$

$$T = \frac{\lambda_E - \lambda_R}{\lambda_E - \lambda_C} = \frac{1 - 0.96}{1} = 0.04$$

$\max[P(\omega_i \mid x)] > T$ is always true. **Therefore, we never reject a pattern**.

This is due to the high value of the reject cost (λ_R =0.96).

Note that if you have two classes the reject threshold T should be higher than 0.5 in order to reject patterns (if T<0.5 then no pattern is rejected). If you have multiple classes, C classes, the reject threshold T should be higher than 1/C (if T< 1/C, then no pattern is rejected).

Example 2 – <u>a proper choice of the cost (loss) matrix</u>

$$\Lambda = \begin{pmatrix} \lambda_R & \lambda_R \\ \lambda_{AA} & \lambda_{AS} \\ \lambda_{SA} & \lambda_{SS} \end{pmatrix} = \begin{pmatrix} \lambda_R & \lambda_R \\ \lambda_C & \lambda_E \\ \lambda_E & \lambda_C \end{pmatrix} = \begin{pmatrix} 0.3 & 0.3 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The cost of each decision is:

$$\lambda = \begin{cases} \lambda_C = 0, Correct\ action \\ \lambda_C = 0.3, reject \\ \lambda_C = 1, error \end{cases}$$

The threshold *T* of the Chow's rule is:

$$T = \frac{\lambda_E - \lambda_R}{\lambda_E - \lambda_C} = \frac{1 - 0.3}{1} = 0.7$$

We classify a pattern as belonging to class ω_i if $\max[P(\omega\_i\,|x)] > T$, we reject it otherwise.

We know that:

P(ω_{SANE})=0.85 , P($\omega_{AFFECTED}$) =0.15

$p(x\,|\,\omega_i) = N(\mu_i, \sigma^2)$; i=1 sane; i=2 affected

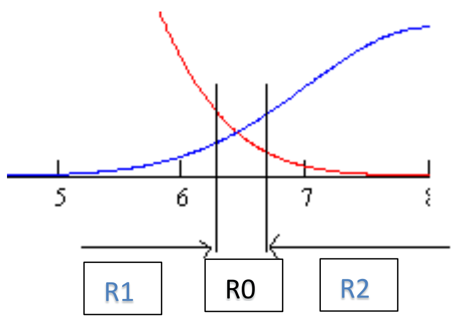
μ_1=4; μ_2=8 ; σ=1

The inequality $\max[P(\omega\_i\,|x)] < T$ defines the region of *rejection* (reject region).

The decision threshold with minimum risk with 0-1 costs is equal to the Bayesian threshold, that we determined in (a): x_b= 6.4337

The Bayes decision rule without rejection is:

$$\max_i[P\ (\omega_i/x)] = \begin{cases} P(\omega_1/x), & x < x_b \\ P(\omega_2/x), & x > x_b \end{cases}$$



If we use the decision region with reject option, the new decision regions are (see the above figure):

$R_1: x \in [-\infty, x_{S1}]$

$R_2: x \in [x_{S2,} \infty]$

$Reject\ region\ RO: x \in [x_{S1}, x_{S2}]$

In the following, we show the calculations to compute $x_{S1}\ and\ x_{S2}$.

$$x_{S1} \rightarrow P(\omega_1/x) < T \text{ and } P(\omega_1|x) > P(\omega_2|x)$$

$$x_{S2} \rightarrow P(\omega_2/x) < T \text{ and } P(\omega_2|x) > P(\omega_1|x)$$

$$x_{S1} \rightarrow P(\omega_1/x) < T \Rightarrow \frac{p(x/\omega_1)P(\omega_1)}{p(x)} < T \Rightarrow \frac{p(x/\omega_1)P(\omega_1)}{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)} < T$$

$$\frac{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)}{p(x/\omega_1)P(\omega_1)} > \frac{1}{T} \Rightarrow$$

$$\Rightarrow 1 + \frac{p(x/\omega_2)}{p(x/\omega_1)}\frac{P(\omega_2)}{P(\omega_1)} > \frac{1}{T} \Rightarrow \frac{p(x/\omega_2)}{p(x/\omega_1)} > \left[\frac{1}{T} - 1\right]\frac{P(\omega_1)}{P(\omega_2)}$$

$\exp[4x - 24] > \left[\frac{1}{T} - 1\right]\frac{P(\omega_1)}{P(\omega_2)} = \left[\frac{1}{0.7} - 1\right]\frac{85}{15} = \frac{3}{7}\frac{85}{15} = \frac{17}{7}$

$4x - 24 > \ln\left(\frac{17}{7}\right); x > 6 + \frac{1}{4}\ln\left(\frac{17}{7}\right);$

$x_{s1} = 6 + \frac{1}{4}\ln\left(\frac{17}{7}\right) \cong 6.2218$

$$x_{S2} \rightarrow P(\omega_2/x) < T \Rightarrow \frac{p(x/\omega_2)P(\omega_2)}{p(x)} < T \Rightarrow$$

$$\Rightarrow \frac{p(x/\omega_2)P(\omega_2)}{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)} < T$$

$$\frac{p(x/\omega_1)P(\omega_1) + p(x/\omega_2)P(\omega_2)}{p(x/\omega_2)P(\omega_2)} > \frac{1}{T} \Rightarrow 1 + \frac{p(x/\omega_1)}{p(x/\omega_2)}\frac{P(\omega_1)}{P(\omega_2)} > \frac{1}{T} \Rightarrow$$

$$\Rightarrow \frac{p(x/\omega_1)}{p(x/\omega_2)} > \left[\frac{1}{T} - 1\right]\frac{P(\omega_2)}{P(\omega_1)}$$

$$\Rightarrow \frac{N(\mu_1, \sigma^2)}{N(\mu_2, \sigma^2)} > \left[\frac{1}{T} - 1\right]\frac{P(\omega_2)}{P(\omega_1)}$$

$$\exp[24-4x] > \left[\frac{1}{T}-1\right]\frac{P(\omega_2)}{P(\omega_1)} = \left[\frac{1}{0.7}-1\right]\frac{15}{85} = \frac{3}{7}\frac{15}{85} = \frac{9}{119}$$

$$24-4x > \ln\left(\frac{9}{119}\right) \ ; x < 6-\frac{1}{4}\ln\left(\frac{9}{119}\right) \ ;$$

$$x_{s2} = 6-\frac{1}{4}\ln\left(\frac{9}{119}\right) \cong 6.6455$$

We recall that the Bayesian error is:

Total Error probability=15.14 10^{-3}

False positive rate=6.35 10^{-3}

False negative rate= 8.79 10^{-3}

Introducing the reject option, we now have:

False positive rate:

$$\int_{x_{S2}}^{\infty} p(x/\omega_{SANE})\, P(\omega_{SANE})\, dx \cong 3.45 \ \times 10^{-3}$$

False negative rate:

$$\int_{-\infty}^{x_{S1}} p(x/\omega_{AFFECTED})\, P\left(\omega_{Affected}\right) dx \cong 5.65 \ \times 10^{-3}$$

Total Error probability = 9.12 10^{-3}

The error is now lower than the Bayesian one without reject option. However, we do NOT classify (we reject) 15 patients every 1000.

$$\int_{x_{S1}}^{x_{S2}} p(x/\omega_{AFFECTED})\, P\left(\omega_{Affected}\right) dx + \int_{x_{S1}}^{x_{S2}} p(x/\omega_{SANE})\, P(\omega_{SANE})\, dx \cong 15.22 \ \times 10^{-3}$$

It is worth noting the difference of the Chow's criterion with respect to the empirical criterion we initially used. In the latter case, we misclassify 7 patients every 1000, we reject 24 patients every 1000.

	Bayesian	Minimum Risk	Reject (empirical thresholds)	Reject (Chow)
Err	15 10^{-3}	29 10^{-3}	7 10^{-3}	9 10^{-3}
False Positive	6 10^{-3}	27 10^{-3}	2.5 10^{-3}	3.5 10^{-3}
False Negative	9 10^{-3}	2 10^{-3}	4.5 10^{-3}	5.5 10^{-3}
Rejected	-	-	24 10^{-3}	15 10^{-3}